Stockpile Reliability Assessment

Nick Hengartner
Statistical Sciences
Los Alamos National Laboratory
Los Alamos, New Mexico
USA
nickh@lanl.gov

Alyson Wilson
Statistical Sciences
Los Alamos National Laboratory
Los Alamos, New Mexico
USA
agw@lanl.gov

Abstract

This paper summarizes some recent advances in semiparametric estimation with surrogate variables that has been motivated by stockpile reliability assessments. The novelty of our results are (1) our estimates are efficient; (2) we can accomodate continuous surrogate and response variables.

1 Introduction

The Los Alamos National Laboratory is entrusted with the evaluation and certification of the safety and reliability of the United States nuclear weapons stockpile. This mission presents unique challenges that is inspiring the development of new statistical methodologies whose common theme is the integration and combination of information from diverse sources (e.g., physics-based computer simulations, historical test data, subsystem tests, accelerated again tests, etc.). This talk will focus on a particular aspect of combining heterogeneous data sources that has been developed for the assessment of munition stockpiles.

2 Problem Statement

Suppose that for each unit in a stockpile, we have the option of performing either of two tests: A destructive test to determine the status Z (pass or fail) of the unit, or, a destructive test measuring the performance X of the unit, whose impact on the system reliability is known to be $\mathbb{P}[Z = 1|X = x] = R(x;\theta)$, with unknown parameter θ. Since both tests are destructive, one never gets to observe both the status and the performance for the same unit. Our problem is to assess the reliability of the stockpile using from independent samples of system status $Z^n = (Z_1, \ldots, Z_n)$ and system performances $X^m = (X_{n+1}, \ldots, X_{n+m})$. This is possible if θ is known. For example, the estimator (with known θ)

$$\hat{p} = \frac{n}{n+m}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) + \frac{m}{n+m}\left(\frac{1}{m}\sum_{i=n+1}^{n+m} R(X_i,\theta)\right)$$

is unbiased for the stockpile reliability and has smaller variance than either $n^{-1}\sum_{i=1}^{n} Z_i$ or $m^{-1}\sum_{j=n+1}^{n+m} R(X_j,\theta)$.

The problem thus reduces to estimating the parameter θ. The parameter θ can not be estimated from independent samples of system status and system performance alone. To link the two samples and make the estimation of θ possible, suppose that both data sets are supplemented with a common covariate measurement W on each unit. This covariate, related to both the performance X and the status Z, is such that it does not offer additional information about the distribution of the status variable Z that is not already contained in the output measurement X. Thus the covariate W acts as a surrogate predictor for X, with the pairs $\{(X_i, W_i), i = n+1, \ldots, n+m\}$ forming a validation sample to precise their relationship. Age and storage conditions are good examples of surrogate variables

A similar data structure also arises in medical application where it is known as the errors-in-covariable with validation sample problem. See for example Wang and Pepe (2000), Pepe and Fleming (1991), Pepe

(1992) and Stefanski and Lee (1995). This problem has also been recently considered in econometrics by Wang et al (2002).

3 Identifiability

Denote by $R(z,x;\theta) = \mathbb{P}[Z = z|X = x]$, $p(z|w) = \mathbb{P}[Z = z|W = w]$, $F(x|w) = \mathbb{P}[X \leq x|W = w]$, and $f(x|w) = \partial F(x|w)/\partial x$. The surrogacy assumptions states that the status Z and the surrogate covariate W are conditionally independent given the performance measure X, that is $\mathbb{P}[Z|X,W] = \mathbb{P}[Z|X]$. When it holds, the conditional probability of the status variable Z given the surrogate W is

$$p(z|w) = \int \mathbb{P}[Z = z|X = x, W = w] f(x|w) dx = \int R(z,w;\theta) f(x|w) dx.$$

Since both $p(z|w)$ and $f(x|w)$ can be estimated from the data sets

$$\mathcal{D}_1 = \{(Z_i, W_i); i = 1, \ldots, n\} \text{ and } \mathcal{D}_2 = \{(X_i, W_i); i = n+1, \ldots, n+m\}, \tag{1}$$

this suggests that the under suitable conditions, the parameter θ may be estimated as well. The need of additional conditions are necessary is seen by considering the case where the performance and the surrogate variables are independent. In that case

$$p(z|w) = \int R(z,x;\theta) f(x|w) dx = \int R(z,x;\theta) f(x) dx = p(z)$$

which does not depend on the surrogate. Depending on the specific distribution of X, the parameter θ may not always be identifiable. The following lemma gives a sufficient condition for identifiability of the reliability function.

LEMMA *If $\{f(\cdot|w)\}$ is a complete family of densities, then the reliability function $R(z,x)$ is identifiable.*

Proof. Suppose to the contrary: there exists $R_1(z,x) \neq R_2(z,x)$ such that

$$0 \quad = \quad p_1(z|w) - p_2(z|w) = \int \{R_1(z,x) - R_2(z,x)\} f(x|w) dx \qquad \forall w.$$

Completeness of the $\{f(\cdot|s)\}$ implies that $R_1(z,x) - R_2(z,x) \equiv 0$, contradicting the hypothesis.

It follows from the Lemma that if the parameter θ is identifiable in the (unobservable) model $R(z,x;\theta)$, then θ will be identifiable in our setting whenever the collection of densities $\{f(\cdot|w)\}$ is complete.

4 Estimation

To gain insight into the proposed estimation method, let us consider first the case of known conditional distribution of X given W. This may be seen as a limiting case in which we have an unlimited sample of pairs (X,W). Let

$$p(z|w,\theta) = \int R(z,x;\theta) f(x|w) dx = \mathbb{E}[R(z,X;\theta)|W = w].$$

The maximum likelihood estimator for θ solves the estimating equation

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p(Z_i|W_i,\theta) = 0.$$

Under mild regularity assumptions to justify interchanging derivatives and integrals, the score function is

$$\begin{aligned} S(z,w;\theta) \quad &= \quad \frac{\partial}{\partial \theta} \log p(z|w,\theta) = \int \frac{\partial}{\partial \theta} R(z|x;\theta) \frac{f(x|w)}{p(z|w;\theta)} dx \\ &= \quad \frac{\mathbb{E}[\frac{\partial}{\partial \theta} R(z|X;\theta)|W = w]}{\mathbb{E}[R(z|X;\theta)|W = w]}, \end{aligned} \tag{2}$$

which is easily verifies to be also the minimizer of

$$\Gamma(\alpha; z, w, \theta) = \mathbb{E}\left[\left(2\frac{\partial}{\partial\theta}\sqrt{R(z|X,\theta)} - \alpha\sqrt{R(z|X,\theta)}\right)^2 \middle| W = w\right]. \tag{3}$$

When the conditional distribution of X given W is not known, Pepe and Reilley (1995) propose to use the validation sample to estimate the score function S by a ratio of conditional averages. Their method requires that both the status Z and the surrogate variable W be discrete. In this paper, we consider an alternative estimate of the score function using nonparametric regression.

Specifically, we propose to estimate the score function $\widehat{S}(z, w; \theta)$ by minimizing with respect to α an empirical counterpart of (3)

$$\sum_{i=n+1}^{n+m} K\left(\frac{w - W_i}{h}\right)\left(2\frac{\partial}{\partial\theta}\sqrt{R(z|X_i,\theta)} - \alpha\sqrt{R(z|X_i,\theta)}\right)^2,$$

where $K(u)$ is a symmetric smoothing of sufficiently high order, and h the bandwidth that goes to zero as the validation sample size m increases.

With the estimate for the score function, we estimate the parameter θ as the solution of

$$\sum_{i=1}^{n} \widehat{S}(Z_i, W_i; \theta) = 0. \tag{4}$$

The following theorem states that this estimator has good statistical properties.

THEOREM *Let $\hat{\theta}$ solve (4). If*

A1 The parameter in θ is identifiable.

A2 Both $n, m \longrightarrow \infty$.

A3 The smoothing kernel $K(\cdot)$ is bounded with compact support and satisfies

- *a $\int K(u)du = 1$*
- *b $K(\cdot)$ is of order r*

A4 The bandwidth h satisfies $\sqrt{nm}h^{2r} \longrightarrow 0$ and $\sqrt{nm}h \longrightarrow \infty$.

A5 The joint density $f(x, w)$ of the joint density of X and W is r times continuously differentiable in w. Let

$$M_h(x, w) = \sup_{|u-w|\leq h}\left|\frac{\partial^r}{\partial w^r} f(x, u)\right|.$$

The functions

$$\begin{aligned} F_1(z, w) &= \frac{\int \partial_\theta R(z, x; \theta) M_h(x, w)dx}{\int R(z, x; \theta) f(x, w)dx} \\ F_2(z, w) &= \left(\frac{\partial}{\partial\theta}\log \mathbb{P}[Z|W, \theta]\right) \cdot \frac{\int R(z|x, \theta) M_h(x, w)dx}{\int R(z|x, \theta) f(x, w)dx} \end{aligned}$$

have finite expectations.

then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta_0))$$

with

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\mathbb{P}[Z|W,\theta]\right)^2\right]$$

the Fisher information matrix.

The proof involves approximating the estimator $\hat{\theta}$ by a suitable U-statistics using common tricks in nonparametric smoothing. The conclusion follows from the CLT for U statistics.

We have several concluding remarks.

1. Under the conditions of the Theorem, estimator $\widehat{\theta}$ is efficient as its asymptotic variance is the same as if we would have known the joint distribution of X and W.

2. The asymptotic distribution of the estimator $\widehat{\theta}$ is not sensitive to the choice of the smoothing parameter h which can be chosen to be in
$$\left(\frac{1}{m}\right)^{1-\varepsilon} \leq h \leq \left(\frac{1}{n}\right)^{\frac{1}{2r}+\varepsilon}$$
for small $\varepsilon > 0$.

3. The size m of the validation sample can be much smaller than n. In particular, the conclusions of the theorem hold provided that
$$m \geq n^{\frac{1}{2r}-\varepsilon}$$
for small $\varepsilon > 0$. Thus smoother joint densities $f(x,w)$ allow for smaller validation samples.

References

[1] Pepe, M.S, (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*, **75**, 237–249.

[2] Pepe, M.S. and Fleming, T.R. (1991). A general nonparametric method for dealing with errors in missing or surrogate covariate data. *JASA* **86**, 108 – 113.

[3] Stephanski, J.H. and Lee, L.F. (1995). Semiparametric estimation of nonlinear errors-in-variables models with validation study. *Journal of nonparametric Statistics* **4**, 365 – 394.

[4] Wang, C.-Y. and Pepe, M.S. (2000). Expected estimating equations to accommodate covariate measurement errors. *Journal of the Royal Statistical Society, Series B*, **62**, 509 – 524.

[5] Wang, Q., Yu, K. and Härdle, W (2002). Likelihood-based kernel estimation in semiparametric errors-in-variables models with validation data. *Technical report, Humboldt Universität zu Berlin.*